PHYSICAL REVIEW E **92**, 040802(R) (2015)

Estimating topological properties of weighted networks from limited information

Giulio Cimini,[1,*] Tiziano Squartini,[1] Andrea Gabrielli,[1,2] and Diego Garlaschelli[3]
[1]*Istituto dei Sistemi Complessi (ISC)-CNR, UoS "Sapienza", Dipartimento di Fisica, Università "Sapienza", Piazzale Aldo Moro 5, 00185 Rome, Italy*
[2]*IMT–Institute for Advanced Studies, Piazza San Ponziano 6, 55100 Lucca, Italy*
[3]*Lorentz Institute for Theoretical Physics, University of Leiden, Niels Bohrweg 2, 9506 Leiden, Netherlands*
(Received 23 September 2014; revised manuscript received 6 February 2015; published 8 October 2015)

A problem typically encountered when studying complex systems is the limitedness of the information available on their topology, which hinders our understanding of their structure and of the dynamical processes taking place on them. A paramount example is provided by financial networks, whose data are privacy protected: Banks publicly disclose only their aggregate exposure towards other banks, keeping individual exposures towards each single bank secret. Yet, the estimation of systemic risk strongly depends on the detailed structure of the interbank network. The resulting challenge is that of using aggregate information to statistically reconstruct a network and correctly predict its higher-order properties. Standard approaches either generate unrealistically dense networks, or fail to reproduce the observed topology by assigning homogeneous link weights. Here, we develop a reconstruction method, based on statistical mechanics concepts, that makes use of the empirical link density in a highly nontrivial way. Technically, our approach consists in the preliminary estimation of node degrees from empirical node strengths and link density, followed by a maximum-entropy inference based on a combination of empirical strengths and estimated degrees. Our method is successfully tested on the international trade network and the interbank money market, and represents a valuable tool for gaining insights on privacy-protected or partially accessible systems.

DOI: 10.1103/PhysRevE.92.040802 PACS number(s): 89.75.Hc, 02.50.−r, 89.65.−s

Reconstructing the statistical properties of a network when only partial information is available represents a key open problem in the field of statistical physics of complex systems [1,2]. Yet, addressing this issue can lead to many concrete applications. A paramount example is provided by financial networks, where nodes represent financial institutions and links stand for the various types of financial ties, such as loans or derivative contracts. These ties result in dependencies among institutions and constitute the ground for the propagation of financial distress across the network. However, due to confidentiality issues, the information that regulators are able to collect on mutual exposures is very limited [3], hindering the analysis of the system resilience to the spreading of financial distress—which depends on the structure of the whole network [4,5]. Typically, the analysis of systemic risk has been pursued by trying to estimate the unknown link weights of the network via a maximum homogeneity principle [6–8], looking for the adjacency matrix with minimal distance from the uniform matrix that also satisfies the imposed constraints (e.g., the budget of individual banks). These approaches are also known as *dense reconstruction* methods, as they assume that the network is fully connected, a hypothesis that represents their strongest limitation. In fact, not only empirical networks do show a very heterogeneous distribution of the connectivity, but such a dense reconstruction leads to systemic risk underestimation [2,8]. More refined methods such as *sparse reconstruction* algorithms [2] allow one to obtain a matrix with an arbitrary level of heterogeneity, however, without prescribing how to identify its proper value; moreover, even when the link density is correctly recovered, systemic risk is again underestimated because of the homogeneity principle used to obtain the link weights. A more recent approach [9,10] instead uses the limited topological information available on the network to generate an ensemble of graphs according to the *configuration model* (CM) [11], where the Lagrange multipliers that define it are replaced by *fitnesses* [12], i.e., node-specific properties assumed to be known—in a way similar to fitness-dependent network models [13]. The estimation of network properties is then carried out within such an ensemble. This method overcomes the limitations of its predecessors, but suffers from another drawback of being usable to reconstruct only binary topologies, whereas systemic risk analysis requires a weighted representation of the network [5].

Here, we aim at overcoming the limitations of these methods and build a procedure to reconstruct weighted networks, resorting on a minimal amount of available information: the total number of connections, and the values of the fitness for each node, whose role will be played by the empirical node strengths. Briefly, our method consists in estimating the number of connections for each node via the standard CM calibrated on the fitnesses, and then in using these values as well as node strengths to assess individual link weights through an *enhanced configuration model* (ECM) [14]. To validate our method, we use two real instances of economic and financial systems for which we have full information (we will be able to assess unambiguously the accuracy of our method in estimating their topological properties). The first one is the World Trade Web (WTW) [15], where nodes represent countries and links stand for trade volumes: The weight w_{ij} of the link between nodes i and j is the total monetary flux (resulting from the import and export) between these countries [16]. The second one is the Electronic Market for Interbank Deposits (E-mid) [17], where nodes represent banks and links stand for loan contracts: w_{ij} is the total amount of liquidity exchanged between banks i and j [18]. In both

*giulio.cimini@roma1.infn.it

cases, the strength of node i is defined as $s_i^* = \sum_j w_{ij}$, while its degree or number of connections is $k_i^* = \sum_j a_{ij}$ (where $a_{ij} := 1 - \delta_{[w_{ij},0]}$).

We now give a detailed explanation of our network reconstruction method, focusing on the main statistical assumptions. Our goal is to find the optimal estimate for $X(G_0)$, the value of a topological property X for a real network G_0, on the basis of the limited available information on G_0 itself: the total number of nodes N and links L, and the whole strength sequence $\{s_i^*\}_{i=1}^N$. Such quantities will act as constraints in the estimation procedure: The idea is to consider G_0 as drawn from the appropriate maximally random ensemble Ω of weighted graphs compatible with such constraints, so that $X(G_0)$ can be estimated as $\langle X\rangle_\Omega$ (the average of X over the ensemble). In other words, we map the problem of evaluating $X(G_0)$ into that of choosing the optimal ensemble Ω compatible with the known constraints. The method proceeds in two main steps, each based on a key assumption.

(I) We first reconstruct the binary topology of G_0. To this end, if we knew the degree k^* for each node of the network, we could use the standard approach of the CM [11,19,20], which consists in generating an ensemble $\Omega_{\rm CM}$ of networks which is maximally random, except for the ensemble average of the node degrees $\{\langle k_i\rangle_{\Omega_{\rm CM}}\}_{i=1}^N$ that are constrained to the empirical values $\{k_i^*\}_{i=1}^N$ observed for G_0. This leads to a probability distribution over $\Omega_{\rm CM}$ of all possible binary graphs, which is defined via a set of Lagrange multipliers $\{x_i\}_{i=1}^N$ (one for each node) associated to the constraints $\langle k_i\rangle_{\Omega_{\rm CM}} \equiv k_i^*\ \forall i$ [21]. Once all $\{x_i\}$ are found, the CM reduces to having a link between nodes i and j with probability [21]

$$p_{ij} = \frac{x_i x_j}{1 + x_i x_j}, \tag{1}$$

independently of all other links. Here, however, we are studying the case where individual node degrees are unknown, yet we know the total number L of links. Thus we cannot directly use the CM; to overcome this drawback, we resort to the *fitness* model [12], which assumes the network topology to be determined by an intrinsic node property called fitness. This approach has been successfully used in the past to model several economic and financial networks, by assuming a connection between fitnesses and Lagrange multipliers [13,17,22]. We thus make the following ansatz: The strengths $\{s_i^*\}_{i=1}^N$ (for which we have full information) are interpreted as node-specific fitnesses, induced by node degrees. In particular, we assume strengths to be linearly proportional to the degree-induced Lagrange multipliers $\{x_i\}_{i=1}^N$ of the CM, with an unknown proportionality constant z: $x_i \equiv \sqrt{z}s_i^*\ \forall i$ (see Fig. 1 and the discussion below). Once z is determined, we can estimate the unknown node degrees as their average values in such strength-induced $\Omega_{\rm CM}$. The first step of our method is thus the estimation of the constant z, which is achieved by equating the average number of links of a graph belonging to $\Omega_{\rm CM}$, computed through Eq. (1) with $x_i = \sqrt{z}s_i^*$, to the (known) total number L of links in G_0:

$$\langle L\rangle_{\Omega_{\rm CM}} \equiv \frac{1}{2}\sum_i \sum_{j(\neq i)} \frac{z s_i^* s_j^*}{1 + z s_i^* s_j^*} = L. \tag{2}$$

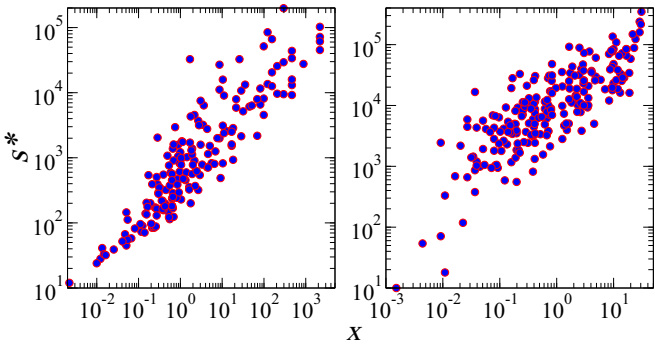

FIG. 1. (Color online) Relation between node strengths $\{s^*\}$ and their degree-induced Lagrange multipliers $\{x\}$ from CM (obtained by knowing the whole degree sequence). The linearity of such a relation is at the basis of the ansatz of our method that $x_i \propto s_i^*\ \forall i$. The left panel refers to WTW, and the right panel to E-mid.

Since $\{s_i^*\}_{i=1}^N$ are known, Eq. (2) is an algebraic equation in z with a single solution for $z > 0$, which is then used to estimate the unknown degrees of G_0:

$$\langle k_i\rangle_{\Omega_{\rm CM}} = \sum_{j(\neq i)} p_{ij} = \sum_{j(\neq i)} \frac{z s_i^* s_j^*}{1 + z s_i^* s_j^*} \quad \forall i. \tag{3}$$

We have thus obtained, for our network G_0, an estimate for node degrees—through Eq. (3)—and for the single link probability p_{ij}—through Eq. (1) with $x_i \equiv \sqrt{z}s_i^*\ \forall i$.

(II) We then reconstruct the weighted topology of G_0. Again, if we had full information on the node degrees, we could use the ECM [14], a more sophisticated version of the CM obtained by constraining the ensemble averages of node degrees $\{\langle k_i\rangle_{\Omega_{\rm ECM}}\}_{i=1}^N$ to $\{k_i^*\}_{i=1}^N$ and node strengths $\{\langle s_i\rangle_{\Omega_{\rm ECM}}\}_{i=1}^N$ to $\{s_i^*\}_{i=1}^N$ [19], and building a maximally random ensemble of weighted graphs compatible with these constraints. The ECM prescribes that two Lagrange multipliers $\{a_i, b_i\}$ are associated to each node i, so that the ensemble probability q_{ij} that any two nodes i and j are connected and the ensemble average weight $\tilde{w}_{ij}$ for such a link become [23]

$$q_{ij} = \frac{a_i a_j b_i b_j}{1 + a_i a_j b_i b_j - b_i b_i}, \quad \tilde{w}_{ij} = \frac{q_{ij}}{1 - b_i b_j}. \tag{4}$$

The problem of missing information on the degree sequence can, however, be overcome owing to the CM-based estimation of step I of our method: We can use the degrees estimated via Eq. (3), together with the empirical node strengths $\{s_i\}_{i=1}^N$, to build $\Omega_{\rm ECM}$ by solving the system of $2N$ nonlinear equations that define it:

$$\begin{aligned} \langle k_i\rangle_{\Omega_{\rm CM}} &= \sum_{j(\neq i)} q_{ij} = \sum_{j(\neq i)} \frac{a_i a_j b_i b_j}{1 + a_i a_j b_i b_j - b_i b_i} \\ s_i^* &= \sum_{j(\neq i)} \tilde{w}_{ij} = \sum_{j(\neq i)} \frac{q_{ij}}{1 - b_i b_j} \quad \forall i. \end{aligned} \tag{5}$$

The solution is the set of Lagrange multipliers $\{a_i, b_i\}_{i=1}^N$ that allow one to obtain the single linking probabilities $\{q_{ij}\}_{i,j=1}^N$ and the average weights $\{\tilde{w}_{ij}\}_{i,j=1}^N$ as of Eq. (4). $X(G_0) = \langle X\rangle_{\Omega_{\rm ECM}}$ can then be computed either analytically, or

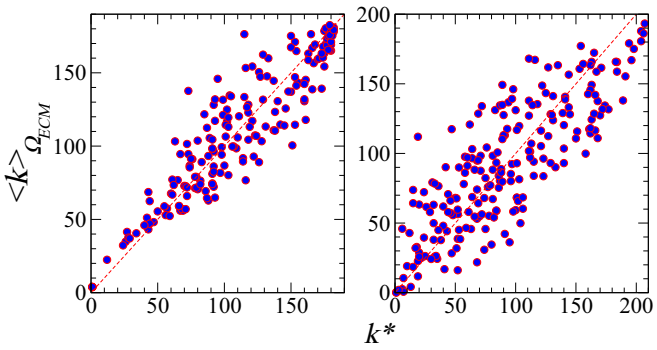

FIG. 2. (Color online) Relation between k^* and $\langle k\rangle_{\Omega_{\rm ECM}}$ for WTW (left panel) and E-mid (right panel).

numerically (on a representative set of networks drawn from $\Omega_{\rm ECM}$).

We now move to validation of the method. First, in order to check whether $\Omega_{\rm ECM}$ defined above is a proper ensemble from which to draw the real network G_0, we compare $\forall i$ the empirical degree k_i^* with $\langle k_i\rangle_{\Omega_{\rm ECM}} = \sum_{j(\neq i)} q_{ij}$ estimated through our method. As Fig. 2 shows, this results in a scattered cloud around the identity, whose behavior reflects the noisy yet very high correlation between these values. We then focus on the topological properties that are commonly regarded as the most significant for describing a weighted network and its binary structure: the average nearest-neighbor strength $s^{\rm nn}$ and the weighted clustering coefficient c^w, and the average nearest-neighbor degree $k^{\rm nn}$ and the binary clustering coefficient c^k [19]. Figure 3 shows a remarkable agreement between these quantities computed on G_0 and their ECM ensemble averages, which can therefore be used as good estimates for the real quantities $X(G_0)$. Such a test reveals the accuracy of our method in reconstructing the topological properties of the real network.

We remark that the applicability of our method strongly depends on the accuracy of the ansatz of whether the CM induced by node strengths is able to provide good estimates for the unknown degrees. The validity of such an ansatz can, however, be assessed through a scatter plot of node strengths versus their degree-induced Lagrange multipliers computed via CM (as shown in Fig. 1), or simply of node strengths versus node degrees (as for small degree values it is $k_i^* \propto x_i$ [19]). In any event, our assumption $x_i \propto s_i\ \forall i$ derives from a simple argument, corroborated by similar evidence found in the analysis of several economic and financial networks [13,17,22]: The more important a node, the bigger we expect its degree to be. This means that any measure of importance of a node is likely to be a monotonic function of the Lagrange multiplier that controls the degree of that node. If such a measure is strictly positive, then its simplest dependency on x is linear. Here, we are expecting the strengths to directly reflect the nodes' importance, with the advantage that, for any network, they always provide a unique proxy—hence we do not need to look for case-specific external quantities.

Another important remark is that our method is based on a combination of CM and ECM rather than directly on the weighted configuration model (WCM) [21], because the latter not only fails to reproduce the network topological properties (as shown by Fig. 3), but it also predicts a far denser network than observed. This happens not because strengths carry a "lower level" information than that of degrees. Rather, they can be used to infer the degrees themselves, and this is what we point out: The information on strength values should not be used to directly reconstruct the network, but to estimate the degree first, and only then to compute the quantities of interest. In this respect, note that using directly the knowledge of the strength sequence and number of links as fixed constraints to build a maximum-entropy ensemble would result in different mathematical expressions. In particular, we would arrive at a variant of Eq. (4) where $a_i = a\ \forall i$ and, just as the WCM, this model gives a bad prediction of the network, leading to the conclusion that inferring the links' presence first is a crucial step of our approach, indispensable to achieve a faithful network reconstruction. The fundamental reason behind this is that, in the absence of topological constraints (i.e., when only the strengths are enforced), the method would assign equal probability to all the configurations that have the same strength sequence. The number of such configurations is extremely large, resulting in an enormous entropy of the network ensemble and in a consequent lack of an accurate reconstruction. By contrast, the specification of degrees constrains the system much more strongly, and results in a definitely smaller entropy [19].

Further work is needed to address several issues that remain open, including testing our method on higher-order topological properties. In this respect, the fact that in our method the probability of generating a graph factorizes into the probabilities of connecting the different pairs of nodes (which in the context of the fitness model is known as the *independence of dyads*) could be seen as making the method inherently inadequate to reproduce, e.g., community structures or spatial dependencies. This inadequacy is probably the cause of the residual deviations between the real networks and our reconstructed ensembles that appear in Fig. 3. We should, however, emphasize that the independence of dyads is not postulated by us at any stage; rather, it emerges naturally from the enforcement of purely local constraints (defined as sums of degrees and/or strengths over neighbors). In fact, when reconstructing a network from such constraints, the maximum-entropy method automatically generates independent dyads as the unbiased solution to the inference problem. Paradoxically, introducing (more realistic) dependencies among dyads would result in a biased inference.

Finally, we remark that in its present version our method exploits very limited information, which is indeed minimal but is also what is often (and only) available for economic and financial systems: Besides global statistics (N and L), the strengths (that can be the operating revenue of firms, or the tier-1 capital of banks) are or should be accessible public data. In conclusion, our method is particularly useful to overcome the lack of topological information that often hampers systemic risk estimation in financial networks. More generally, our method can be applied to any complex system for which the information on the dependencies among its components is limited (because of observational limitations and difficulties in collecting data). Yet, one should always keep in mind the limitations of maximum-entropy approaches based on local constraints in reconstructing high-order properties of inherently nonrandom systems (such as some biological and

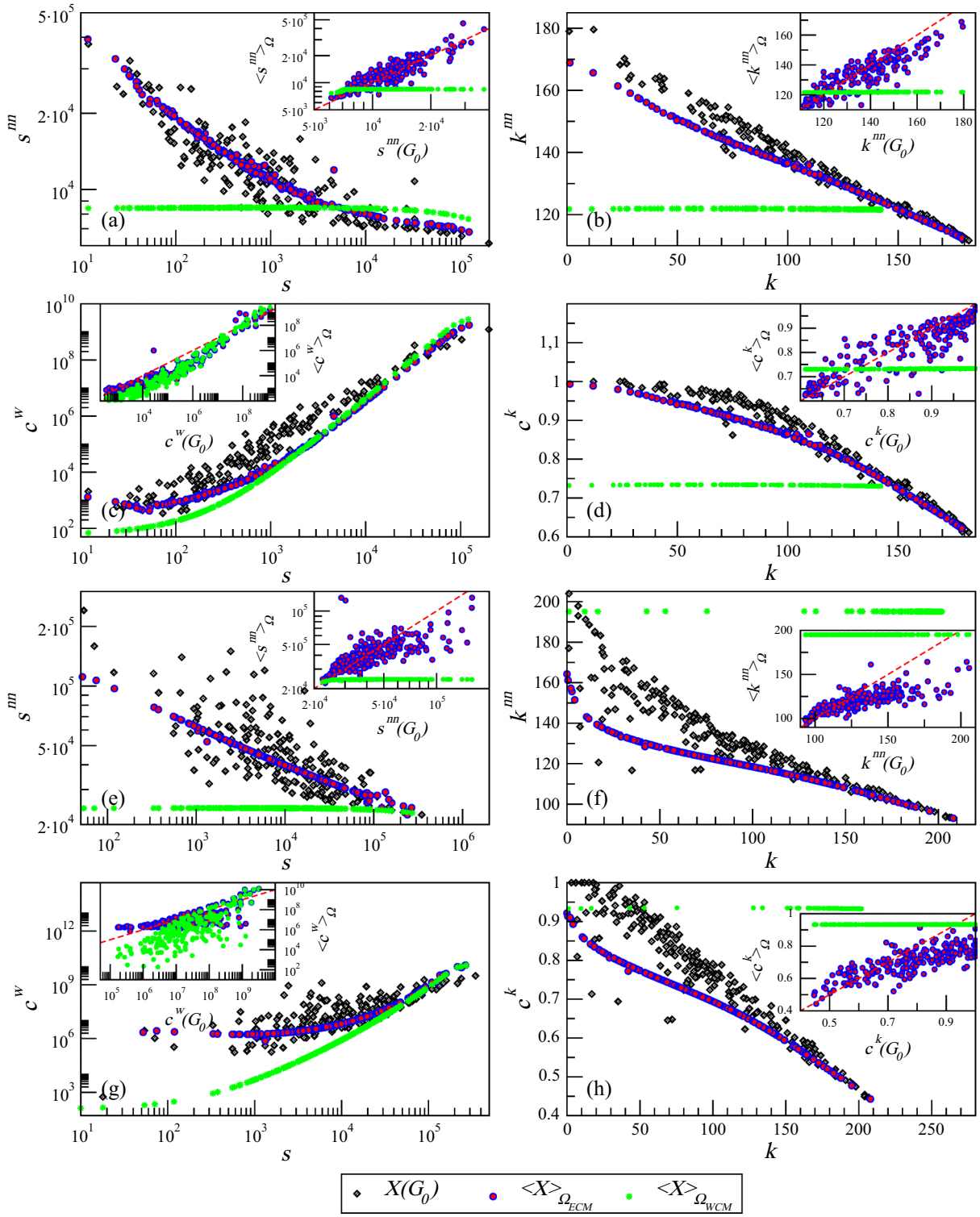

FIG. 3. (Color online) Scatter plots of (a), (e) s vs s^{nn}, (b), (f) k vs k^{nn}, (c), (g) s vs c^{w}, and (d), (h) k vs c^{k} for the real quantities $[X(G_0)]$, those estimated by our method ($\langle X\rangle_{\Omega_{\rm ECM}}$), and those computed by WCM-based reconstruction ($\langle X\rangle_{\Omega_{\rm WCM}}$). Insets: Relations $X(G_0)$ vs $\langle X\rangle_{\Omega_{\rm ECM}}$ and $X(G_0)$ vs $\langle X\rangle_{\Omega_{\rm WCM}}$ for the same quantities. Upper plots (a)–(d) refer to WTW, and lower plots (e)–(h) to E-mid.

technological networks), but still can provide useful insights on local features [19].

This work was supported by the European Union 7th Framework Program—projects GROWTHCOM (Grant No. 611272) and MULTIPLEX (Grant No. 317532), the Italian PNR project CRISIS-Lab and the Netherlands Organization for Scientific Research (NWO/OCW). D.G. acknowledges support from the Dutch Econophysics Foundation (Stichting Econophysics, Leiden, the Netherlands) with funds from beneficiaries of Duyfken Trading Knowledge BV (Amsterdam, the Netherlands).

[1] A. Clauset, C. Moore, and M. E. J. Newman, Nature (London) **453**, 98 (2008).
[2] I. Mastromatteo, E. Zarinelli, and M. Marsili, J. Stat. Mech. (2012) P03011.
[3] S. Wells, Bank of England's Working Paper No. 230, 2004, http://www.bankofengland.co.uk/archive/Documents/historicpubs/workingpapers/2004/wp230.pdf.
[4] S. Battiston, D. Gatti, M. Gallegati, B. Greenwald, and J. Stiglitz, J. Econ. Dyn. Control **36**, 1121 (2012).
[5] S. Battiston, M. Puliga, R. Kaushik, P. Tasca, and G. Caldarelli, Sci. Rep. **2**, 541 (2012).
[6] I. van Lelyveld and F. Liedorp, Int. J. Cent. Bank. **2**(2), 99 (2006).
[7] H. Degryse and G. Nguyen, Int. J. Cent. Bank. **3**(2), 123 (2007).
[8] P. Mistrulli, J. Bank. Fin. **35**, 1114 (2011).
[9] N. Musmeci, S. Battiston, G. Caldarelli, M. Puliga, and A. Gabrielli, J. Stat. Phys. **151**, 720 (2013).
[10] G. Caldarelli, A. Chessa, A. Gabrielli, F. Pammolli, and M. Puliga, Nat. Phys. **9**, 125 (2013).
[11] J. Park and M. E. J. Newman, Phys. Rev. E **70**, 066117 (2004).
[12] G. Caldarelli, A. Capocci, P. De Los Rios, and M. A. Muñoz, Phys. Rev. Lett. **89**, 258702 (2002).
[13] D. Garlaschelli and M. I. Loffredo, Phys. Rev. Lett. **93**, 188701 (2004).
[14] R. Mastrandrea, T. Squartini, G. Fagiolo, and D. Garlaschelli, New J. Phys. **16**, 043022 (2014).
[15] K. S. Gleditsch, J. Conflict Resolut. **46**, 712 (2002).
[16] We use trade volume data for year 2000.
[17] G. De Masi, G. Iori, and G. Caldarelli, Phys. Rev. E **74**, 066112 (2006).
[18] We consider loans for year 1999 (aggregated on annual scale) [17].
[19] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.92.040802 for details of the method, tests, and a sample application on biological networks.
[20] S. Dorogovtsev, *Lectures on Complex Networks* (Oxford University Press, 2010).
[21] T. Squartini and D. Garlaschelli, New J. Phys. **13**, 083001 (2011).
[22] D. Garlaschelli, S. Battiston, M. Castri, V. Servedio, and G. Caldarelli, Physica A **350**, 491 (2005).
[23] D. Garlaschelli and M. I. Loffredo, Phys. Rev. Lett. **102**, 038701 (2009).